DOCUMENT RESUME

ED 122 629 FL 007 668

AUTHOR Swain, Merrill
TITLE Evaluation of Bilingual Education Programs: Problems and Some Solutions.
PUB DATE Feb 76
NOTE 22p.; Paper presented at the conference of the Comparative and International Education Society (Toronto, Ontario, February 25-27, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Bilingual Education; Bilingual Schools; Bilingual Students; Bilingual Teachers; Educational Assessment; English; *Evaluation Methods; French; *Program Evaluation; Research Design; *Research Problems; Second Language Learning; *Statistical Analysis; Testing; *Testing Problems; Test Interpretation

ABSTRACT

One of the overriding problems of any psychometric study is that aspects of behavior that are nonquantifiable or difficult to quantify are not considered. The other side of this coin is that those aspects of behavior that are easiest to quantify tend to be what are examined. In evaluating bilingual education programs, problems are found in the research design, the students tested, the nature of the tests used, the methods of analysis employed, and the interpretation of findings. Research design has usually involved comparison of bilingual program students with students in a traditional program. This is easier than analyzing the goals and objectives of the program and what a student should achieve, but not always accurate. Selection of a control group is difficult because: (1) the characteristics of bilingual education students are not always known; (2) significant characteristics may be difficult to quantify; and (3) some characteristics may not be present among students in regular programs. There are three main problems in test selection: (1) neither standardized tests nor specially developed tests give accurate comparisons between the special and regular program students; (2) the content measured in tests is limited, and their data are used too extensively; and (3) tests tend to be the only accepted means of obtaining performance data. Statistical comparisons are needed but are overused. The methods of analysis most commonly used do not take into account possible interrelationships between separate test scores in a battery. Misunderstanding of statistics and, most importantly, a lack of information about the program being evaluated are two problems with the interpretation of findings. Many evaluation problems can be overcome by supplementing the data with case studies of students, observational data, and specific information about program structure, progress and content. (CHK)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED122629

EVALUATION OF BILINGUAL EDUCATION PROGRAMS: PROBLEMS AND SOME SOLUTIONS

Merrill Swain

Ontario Institute for Studies in Education
252 Bloor Street West
Toronto, Ontario

FL007668

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at the Conference of the Comparative and International Education Society in Toronto, February 25-27, 1976.

Evaluation of Bilingual Education Programs: Problems and Some Solutions

Merrill Swain
Ontario Institute for Studies in Education

I have been asked to address myself in this paper to the problems of evaluating bilingual education programs under those conditions where the independent variable in the research design is the bilingual education program itself. In other words, I will be considering situations where the main concern of the evaluator is to determine the effects on students (dependent variables) of a bilingual education program (independent variable).

Although I have been asked to discuss the problems associated with research of this sort, I want to make it clear from the outset that, in spite of the problems, I consider the research to be of considerable value within the educational context in which it has been undertaken. It can provide essential feedback to the programs themselves often resulting in program change. Furthermore, such evaluations have provided individuals considering the implementation of a bilingual education program with information about the results of a variety of options, one of which may be applicable to their community given their own particular needs and characteristics. And finally, although a single evaluation carries little generalizability in and out of itself, "such case studies are necessary if we are to begin to develop a theory of

---

[1]I would like to thank Dr. L.D. McLean and Dr. H.C. Barik who took the time to read and comment on an earlier draft of this paper, and with whom I have had many discussions about the evaluation of bilingual education programs.

bilingual education which will enable us to generalize the evidence from the individual studies and to account for their often contradictory findings" (Paulston, 1975, p. 5).

In discussing the problems of evaluating bilingual education programs, I want to consider separately the following related aspects: (1) the research design; (2) the students tested; (3) the nature of the tests used; (4) the methods of analysis employed; and (5) the interpretation of findings. But let me introduce the central problem by means of an anecdote.

Last spring I was visited by a professor of linguistics who is well-known for her contributions to the fields of second language teaching, language planning, and bilingual education. One of her major priorities while in the Toronto area was to visit several of the experimental programs in bilingual education that I had been involved with. (For evaluations of these programs, see Barik & Swain, 1976; in press.)

One of these programs, referred to as an "early French immersion program", begins at the kindergarten level. The pupils entering the program are, with few exceptions, from English speaking homes and have had little or no exposure to the French language. In spite of this, beginning with the first day of school, the teacher speaks to them only in French. Although typically a native speaker of French, the teacher is bilingual, and therefore responds appropriately - but in French - to the children's incessant questions, requests, stories, and general chatter in English. French continues to be the only language of instruction through grades 1, 2 or 3, at which time a daily period of instruction in English Language Arts - taught in English - is introduced into the

curriculum. The plan in the schools I am involved with is that with each successive year of schooling more time will be devoted to instruction in English until, by around grade 5 or 6, approximately half of the curriculum will be taught using English as the medium of instruction and half using French as the medium of instruction.

Another program which my visitor wanted to see was one which has sometimes been referred to as a "late" French immersion program - in order to contrast it with the "early" immersion program which begins at the primary level. The late immersion program is one which begins around grade 7 or 8. In some programs, for example those in Montreal, students have had instruction in French as a second language (FSL) since the primary grades. In other programs, preceding FSL instruction has been much shorter. The grade 8 classes my visitor went to see had approximately 70% of their curriculum taught to them in French, the rest in English throughout the grade 8 year. Prior to this immersion experience, they had had only one year of FSL instruction - 20 minute periods daily, intensified to 60 minute periods during the last two months of their grade 7 year. Following the immersion year, these students have the option of being able to take history and geography in French, as well as an enriched FSL course.

My visitor wanted to see these programs for at least two reasons. First, she had read most, if not all, of the existing literature concerned with the Canadian immersion programs,[1] and was looking forward to matching reality with the "image" she had constructed of the programs based on her reading of the research reports. Secondly, she had her

---

[1] A bibliography of the literature pertaining to research on immersion education for the majority language child appears in Swain & Bruck (1976).

own biases: she was convinced although she had no confirmation from the immersion research literature, that the late immersion program would produce better results; that is, the students would be "better" in French. I too had my biases - developed from numerous visits to both programs - that the early immersion program would produce better results.

After one day of observation in the grade 1 classes of the early immersion program and the grade 8 classes of the late immersion program, she had changed her mind. She thought that the grade 1 immersion students were more at ease in speaking French than the grade 8 students. They spoke French in the classroom unselfconsciously, with little concern about the grammatical correctness of their speech. They were more fluent and they sounded better. They were clearly enjoying the activities they were involved in and the use of French in the school context seemed natural and unimposed.

That ends my story. What does it tell us about the problems of evaluating bilingual education programs? For me, it brings to the forefront one of the overriding problems of any psychometric study - aspects of behaviour which are non-quantifiable or difficult to quantify are not considered. Judgements of performance - like those just made about the early immersion children's use of French - rarely appear in evaluation reports because they are considered to be too subjective and biased to have any scientific value.

It would be possible to objectify the judgements by having, for example, a number of randomly selected individuals observe the students. It might also be possible to quantify the judgements by having these individuals rate on a 5 or 7 point scale the linguistic behaviour of the students

along a number of dimensions like, for example, ease of talking. If we found that the early immersion pupils were rated higher we would have spent thousands of dollars demonstrating what we already know based on informed opinion. If, however, we found that early immersion pupils were rated lower, we would likely spend considerable time trying to find out what went wrong. Perhaps the dimensions used were poorly worded. Perhaps the raters had lower expectations for adolescents than children, their ratings thus reflecting surprise and pleasure that adolescents were doing so well relative to their expectations.

The other side of the coin to the problem that non-quantifiable aspects of behaviour tend not to be considered is, that those aspects of behaviour that are easiest to quantify tend to be what are examined. This fact, as we shall see, underlies many of the problems associated with problems of evaluating bilingual programs.

i) Research Design. An evaluation of a bilingual education program is typically set up with the following question in mind: "how do the students in the bilingual education program perform relative to how they would be doing if they were not in the program, that is, if they were receiving the 'traditional' or 'usual' form of education?" or, as we will see below "how do the students in the bilingual education program perform relative to students receiving the usual form of education, or relative to students in another type of bilingual program?" In other words, the research design has typically involved the comparison of one group of students to another.

Sometimes I think we have fallen into using this research design by default. That is to say, just like we measure those things which

are easiest to measure, we also use the research design which is easiest to use. It is much easier to ask "is X doing better than Y?" than "how well should X be doing?" The latter question implies knowing what the goals and objectives of the program are, knowing what one wants the students to learn, knowing what minimal level of skills and knowledge acquisition is considered acceptable and developing tests to measure it.

For example, one can ask if students in early immersion programs learn more French than those in late immersion programs? Answering this would involve giving the same battery of tests to both groups of students and determining which group scores better. (This ignores the serious problems involved in giving the same battery of language tests to students from such different age groups (Swain, 1974)). Or one can ask if students in the early immersion program can understand specific vocabulary, grammatical structures, descriptions, etc., given in French that their teachers agree their students should know. To determine this, one needs no comparison group. But one needs explicit statements of program goals and curriculum content. These are often non-existent. Typically then, the evaluator "reverts" to a comparative research design rather than attempt to detail the goals and curriculum content.

I am not suggesting that evaluations using a comparison group are inappropriate. Rather, I am suggesting that the evaluation of a program need not necessarily involve the use of a comparison group -- the selection of which is often a problem as is discussed below. The use of a comparison group should be dependent upon the nature of the questions asked. However, the tendency to ask those questions which are easiest to answer sometimes overrides the tendency to ask the

appropriate questions.

ii) The Students Tested. The main problem associated with the students tested is in deciding whom to test. Educational research is rarely like "laboratory" research where the experimenter can randomly assign children to different groups - in this case, different educational programs. Instead, the students are "assigned" on the basis of numerous factors which often operate outside of the educational system, which are not already readily identifiable, which are often difficult to quantify and which rarely operate randomly in distributing students into educational programs.

Let us consider what this means relative to the students enrolled in the two French immersion programs discussed earlier. In both cases, the programs are optional. In effect this means that for the early immersion children, the parents make the decision whether to enrol their child in the French or English program. For the late immersion students, it is ultimately the student himself who makes the decision, with a little help from his friends, parents and teachers. We know very little about the late immersion student except that typically he has obtained above average grades in his FSL course(s), and/or his friends have, and/or his parents think it is a good idea for him to learn French. In other words, the only reliable characteristic the late immersion student brings to the program is a motivation to continue to learn French.

We also know very little about the early immersion student. We do know, however, a few things about the parents of the early immersion children, or to be more accurate, we know something about the parents in the Ottawa area who enrolled their children in the French immersion program in its first year of operation. For example, when asked why

they enrolled their child in the program they did, French immersion parents frequently indicated that it was because they wanted their child to be bilingual. They did not seem concerned about whether or not the program would be too difficult for their child in that they rarely mentioned it as a worry. However, many parents who enrolled their child in the English program expressed concern that the French immersion program might be an upsetting or confusing experience and/or that the development of basic English skills might suffer. Several parents were certain that their child would not be able to cope in the French immersion program because of special learning or personality problems (Swain, Wesche, & Machin, 1972).

Now, considering the question of whom to test, it goes without saying that the students in the program, or a sample thereof, are to be tested. The problems arise in selecting a control group. A control group is needed when the basic question is "how do students in bilingual education programs perform relative to how they would be doing if they were receiving the usual form of education?" Thus what is needed is a group of students, similar in all respects except that they are enrolled in the regular English program. The preceding discussion should make it apparent, however, that it is rather unlikely that the French immersion students and the English program students will be similar in all respects.

If <u>all</u> the characteristics of the bilingual education group were identifiable and quantifiable, then presumably one could hunt through the population of students in the regular school program and perhaps find a group of students with equivalent characteristics. But the problem is just that - the characteristics of the bilingual education group are not always readily identifiable, nor quantifiable. For

example, in the case of the primary level program the fact that parents of the immersion children do not seem particularly concerned about their childrens' ability to cope with the immersion class may mean that the children in the program are brighter, or more extroverted, or more verbal, or simply more able to cope.

To summarize then, a major problem in evaluating bilingual education programs is in the selection of a control group. Typically a control group cannot be selected because (1) the characteristics of the bilingual education students are not always known; (2) the characteristics that are judged significant may be difficult to quantify; and (3) at least some of the characteristics may not be present among the students in the regular program.

This is a problem without a solution. However, the basic research question can be changed from "how do students in the bilingual education program perform relative to how they would be doing if they were receiving the usual form of education?" to "how do the students in the bilingual education program perform relative to students receiving the usual form of education?" *The question posed this way recognizes that students who participate in the program have characteristics different from those not participating in the program, and that that is part of the very nature of the program.* In other words, it assumes that the particular bilingual education program is one which attracts students who, for example, in the case of the late immersion students are motivated to learn French, or in the case of the early immersion students are more verbal, able to cope, or whatever.

In selecting such a comparison group, however, it is essential that every attempt be made to control variables which are known to affect achievement in school, for example, socioeconomic status and IQ. If the immersion program is attended by working class children, then the comparison group should also consist of children from working class homes.

Caution must be exercised in the interpretation of results when a comparison group is used. One cannot say that the bilingually educated students are doing as well, or better, or worse than they would be doing if they were being schooled in a regular English program, but one can say that they are doing as well, or better, or worse than other students enrolled in the regular English program. Of course, none of these problems arise if the questions asked are non-comparative in nature, relating to attainment of agreed-upon objectives.

If one assumes that the factors related to the distribution of students into educational programs does not change over time, then the results can be generalized to succeeding cohorts. However, the extent to which they change is the extent to which the results are non-generalizable within the operation of the program itself. For example, the fact that immersion research results are released to parent groups means that many who had concerns about the program may change their minds and enter their child in the program. This means that a different population of students -- one that is perhaps more heterogeneous may be enrolled in the immersion program. In such a situation a new evaluation is probably called for to see to what extent the results based on the new population of immersion students replicates that found with the initial group of immersion students.

iii) The Nature of the Tests Used. Let us turn now to consider in more depth the problems associated with the tests used. The first question an evaluator must face is whether to use standardized tests or tests developed with the particular bilingual education program curriculum and goals in mind. There are problems inherent in both.

One advantage of standardized tests is that they are available! The development of a valid and reliable test is an extremely time-consuming process. When we began our evaluation of the early immersion program and decided that we wanted to measure French comprehension, we quickly discovered that there were no tests commercially available that were appropriate either conceptually or linguistically. So we began to develop our own tests of French comprehension (Barik, 1975; in press). From the construction of the first pool of test items through to consultation with teachers in the program, to administration of the test, to revision of test items based on item analysis, to readministration, to final revision, and lastly to obtain norms, took a period of approximately four years. And during that four year period when we were developing the test, we were also using the test to measure student performance - a somewhat unsatisfactory procedure.

One problem associated with the use of standardized test is that some items may be inappropriate. For example, a number of items on achievement and IQ tests standardized in the United States include aspects of American life with which Canadian children may be unfamiliar. This problem becomes more severe the more culturally or linguistically different the group being tested is from the norming population. This makes the use of standardized tests in cross-cultural research extremely problematic.

Another problem associated with standardized tests is that they tend to maximize differences among students in a class while minimizing the differences among classes. As Macnamara (1975) points out, one reason for this is the choice of questions used in the test.

> Questions which deal with matters that have not been dealt with in all or nearly all classes are dropped straight away. Thus, standardized tests give little credit to the inventive or adventurous teacher or programme.
>
> (Macnamara, 1975, p. 51)

An alternative to giving standardized tests is to give tests developed with the curriculum and goals of the particular bilingual education program in mind. Done properly, this is a satisfactory procedure. But doing it properly is not always easy. As has already been noted above, the goals of a program rarely exist in a detailed written form. Developing a satisfactory test means getting agreement among program organizers, teachers and consultants concerning the objectives of the program, and the appropriateness and relevancy of the test items to program goals and curriculum. Furthermore, if there are a great many classes in the program, then the same problem as the standardized test developers face -- that of finding the common content -- is equally a problem here.

Perhaps the most serious problem of using locally developed tests is that it makes comparisons across alternative bilingual education programs difficult if not impossible to make. For example, in the Ottawa area alone there exist several variations of the early immersion theme depending on when the English Language Arts portion of the curriculum is introduced, what proportion of time is devoted to instruction in French and the extent to which French is taught as a subject. Using the same test to measure, for example, French comprehension permits one to make comparisons across these variations. This

is especially important when a program is being developed and variations within the same community context are being tried out, or when the same program is being tried out in different community contexts.

Because most bilingual education programs are either trying to teach a second language or improve an existing one, their evaluations have, not surprisingly, focussed largely on language-related skills. Achievement tests which measure word knowledge, spelling, punctuation, capitalization, grammar and reading comprehension in both languages are the standard fare. These are pencil and paper tests which can be administered to a number of students simultaneously. And they can be scored semi-automatically from answer keys, or by the computer. Tests which must be administered individually or which involve time and judgement in scoring (e.g., tests of creativity) are much less widely used. In other words, what are measured most frequently are those aspects of language that are easiest to measure. What are rarely measured are those aspects of language which are difficult to measure. They may be difficult to measure because they are not well enough understood to develop a relevant test, or because the collection and analysis of the data are simply too time consuming. For example, we had French immersion students at the grade 3 level write short stories for us in French and in English. We also had a comparison group of grade 3 children in the English program write short stories in English. A research assistant spent the better part of several months counting and classifying the words written, the sentences written, the morphological errors, the syntactic errors, the spelling errors, the punctuation errors and the capitalization errors. The number of stories with titles were counted, as were the number of stories which contained dialogue, which were logically sequenced, and so forth.[1] I do not doubt the value of having

---

[1]The results of this study are reported in Swain (1975).

done this - personal comments from teachers and program developers have convinced me of the usefulness from the point of view of deciding where teaching efforts should be concentrated and what teaching aid materials need to be prepared.

But one thing was lacking when we had finished all this counting and classifying, and that was any overall feeling about the stories the children had written. Or to be more accurate - the feeling was there, but it was not to be found in the number and percentages reported. This year we are trying to capture that feeling by asking individuals experienced in teaching children of this age to rate each story on the basis of one quick reading of it. "Please give us your overall impression", we have asked. We will also have individuals make an overall judgment about the stories the children tell, about the fluency with which they speak, about the acceptability of their second language to native speakers, and other such subjective judgments.

In summary then, the problem with tests used in bilingual education evaluations is three-fold: (1) type - standardized tests permit comparisons across programs but reference to test norms are inappropriate to the extent that the groups tested differ from the norming population. Standardized tests tend not to measure learning resulting from innovative curricular practices. Tests developed specifically to measure the outcome of specific curricular inputs, although difficult to develop, are appropriate when group comparisons are not being made. When group comparisons are made, however, the results will obviously be biased in favour of the group receiving the specific curricular input. (2) content - the content measured in tests

is limited. For example, language tests measure the skill aspects of language rather than the communicative, creative or aesthetic aspects. (3) utilization - tests tend to be the only accepted means of obtaining performance data. Test data should be complemented by other data which may well be of a more subjective, non-quantifiable nature.

Methods of analysis: Considering that the basic research question is typically concerned with how the students in the bilingual education program are performing relative to students in the regular program, it is not surprising that the t-test and analysis of variance, which permit the comparison of two or more groups of students on a single measure of performance, are the most commonly used methods of analysis. Slightly more "sophisticated" is the use of analysis of covariance which statistically adjusts for initial differences between the groups. Of course, the dimensions along which the groups differ initially must be identifiable and quantifiable in order to make use of them in the analysis -- and as already noted above, it may be rare that the relevant dimensions are both identifiable and quantifiable.

These methods of analyzing group differences do not take account of interrelationships that might exist between separate test scores in a complete test battery. Multivariate analyses, for example factor analysis and multivariate analysis of variance, provide an alternative approach to data analysis which exploits rather than ignores the fact of test measure interrelationships.

There are at least two reasons why these interrelationships should be examined:

(1) much of a test battery may be redundant. In other words if the results of several tests are highly correlated, then there is no need to administer all the tests even if they purport to measure different things. (2) the interrelationships may have important theoretical implications. In our evaluations we have given a number of tests to measure aspects of language skills in both French and English. Furthermore we presently have data related to language skills for the same children over a period of five years. I want now to examine these data with the general question in mind "what is the relationship between a student's ability in his first language with that in his second language?". More specifically this entails questions like "if a child reads well in English, does he also read well in French?", "if a child is an excellent speller in English, does he also spell well in French?" and "is a child who is verbally glib in English come to be verbally glib in French?". In order to answer these questions, it will probably be important to examine the data at the item level.

In and of themselves, statistical comparisons are needed since they provide critical information about differences between groups. The problems associated with the use of statistics lie in their overuse, in the information that is lost, and in their potential misinterpretation by those unfamiilar with statistics and statistical significance.

The problem of overuse involves three aspects. In the first place, by overuse I mean the use of statistical means of comparison to the exclusion of any other means. What other means am I referring to? Common sense is one. If native speakers of French report that they had no difficulty understanding children in a grade 1 French immersion class and found the experience enjoyable, but struggled to understand students in a grade 8 immersion class and felt uncomfortable because the students were ill at ease, then it will take some

fairly impressive statistics to alter this impression or to improve my understanding of the situation. My understanding would improve by exploring such questions as: are younger children generally easier to talk to than adolescents? do adults have higher, or different, expectations for adolescents than for children with respect to second language abilities? are the difficulties in comprehension related to degree of hesitancy in speech, accent, lack of lexical knowledge, morphological errors, or syntactic errors? We are not yet able to quantify these concepts well enough for statistics to play a major part in our data analysis.

The second type of overuse is application to vast quantities of data. One study[1] I am aware of reports as conclusions a statement like, of the 100 scores associated with achievement, 19 of these favour the bilingual group, 20 favour the unilingual group and 61 show no differences between the groups. My question is "so what"? With that many scores, some of the noted differences are to be expected by chance. Furthermore, the scores where differences were found may be highly interrelated. Without considering their interrelationships, a statement such as the above is essentially meaningless.

The third way in which I mean overuse is in its application to data with the expectation that sense will come out of nonsense. Statistics are sometimes used like a blessing, that when applied to the data will make it holy. If, for example, items on an attitude questionnaire are conceptually or linguistically too advanced for the age level of the children being tested, then no amount of massaging the data (with, e.g., factor analysis) will make

---

[1] presently a confidential document.

the scores obtained make sense. Probably, we have all examined items constituting factors resulting from a factor analysis where the factors made little sense but they were given names anyhow, thus making sense prevail over the nonsense of the data.

A great deal of information is lost when data are submitted to statistical analyses. On the one hand, of course, statistics are useful for precisely the summarizing function they perform. On the other hand, summaries mean that individual differences among students are often ignored, and that details which might generate interesting hypotheses often go undetected. One way of dealing with this problem would be to supplement the statistical data with several case studies of individual students (see, for example Wesche, 1974, Naiman, Frohlich & Stern, 1975).

There will always remain the problem of misinterpretation of reported results by individuals who are unfamiliar with statistics and who do not understand the concept of statistical significance. It probably still happens that otherwise intelligent people consider statistically significant differences and "educationally significant differences" as one and the same thing. Clearly, a distinction must be made between the two, and it is incumbent upon the evaluator to inform educators and parents that statistical differences can provide no more than one source of input into decisions about the educational importance of a noted difference.

The interpretation of findings: There are other problems associated with the interpretation of results than the one just mentioned above. I want to focus on only one other one here -- one that I consider to be the major problem -- a lack of information about the program being evaluated. It is impossible to locate the cause of differing results of programs if one knows nothing about what goes on in the classroom.

This statement appears so obvious as to be absurd. Yet, time and time again I learn of evaluations being undertaken in which the evaluator's only contact with the classes is during the test administration period, if then. Interpretations of evaluation results must rely on information about the programs, including information about curriculum materials used and teaching training and preparation, gathered from such sources as observation in classrooms and interviews with teachers and administrators.

If immersion students do not, for example, spell as well in English as their English-educated counterparts, why should this be so? Is it because less time is spent on spelling drill per se in the bilingual education program? or because different texts are used? or different teaching methods used? etc. Unless it is known that differences do not exist between programs along these dimensions, further speculation about such causal factors as for example, interference from French to English are inappropriate (although a detailed examination of the spelling errors made by students rather than the summary statistic would provide data relevant to such an interpretation).

In conclusion, it would appear that many of the problems associated with the evaluation of bilingual education programs can be overcome by supplementing the psychometric data with observational data from informed observers, case studies of individual students, and specific information about program structure, progress process, and program content. Psychometric data alone did not and could not address the issue of differences between the early and the late immersion program. Inclusion of other sorts of data into evaluation reports might have precluded, however, the marvellous discussions my visitor and I had about the real differences in program outcomes.

References

Barik, H.C. French Comprehension Test, Level 1, Toronto: Ontario Institute for Studies in Education, 1975.

Barik, H.C. French Comprehension Test, Primer, Toronto: Ontario Institute for Studies in Education, in press.

Barik, H.C., & Swain, M. Primary grade French immersion in a unilingual English Canadian setting: The Toronto Study through Grade Two. Canadian Journal of Education, 1976, 1, pp. 39-58.

Barik, H.C., & Swain, M. A Canadian experiment in bilingual education at the grade eight and nine levels: The Peel Study. Foreign Language Annals, in press.

Naiman, N., Frohlich, M., & Stern, H.H. The Good Language Learner, Toronto: Ontario Institute for Studies in Education, 1975. Report submitted to the Secretary of State, Language Programs Branch, 1975.

Paulston, C.B. Ethnic relations and bilingual education: Accounting for contradictory data. Working Papers on Bilingualism, Ontario Institute for Studies in Education, 1975, 6, pp. 1-44.

Swain, M. French immersion programs across Canada: Research findings. Canadian Modern Language Review, 1974, 31, pp. 117-129.

Swain, M. Writing skills of grade three French immersion pupils. Working Papers on Bilingualism, 1975, 7, pp. 1-38.

Swain, M., & Bruck, M. (eds.) Proceedings of the conference on immersion education for the majority child. Canadian Modern Language Review/ La Revue canadienne des langues vivantes, 1976, 32, entire issue 4.

Swain, M., Wesche, M., & Machin, J. French immersion Kindergarten program in the capital area schools: Analysis of interviews with parents and pupils. Mimeo, Modern Language Centre, 1972.

Wesche, M. The Good Adult Language Learner: A study of learning strategies and personality factors in an intensive course. Unpublished doctoral thesis, University of Toronto, 1975.